Proceedings of the International Conference , "Computational Systems for Health & Sustainability"
17-18, April, 2015 - by R.V.College of Engineering,
Bangalore,Karnataka,PIN-560059,INDIA

DISTANCE BASED MINING VIA WATERMARKING

Deepika J
M.Tech Student
Computer Science & Engineering
Srinivas Institute of Technology, Valachil
Mangalore, Karnataka, India

Smeja K
Assistant Professor
Department of Computer Science
Srinivas Institute of Technology
Mangalore, Karnataka, India

Abstract - Protection of the intellectual property is a most important to Provide security and it also provides mechanisms for establishing the ownership of a dataset consisting of multiple objects. Algorithms are used to preserve important properties of the dataset, it is important for data mining operations. Hence for Preservation of Dataset Watermarking Scheme is used and it may distort the original distance graph of the given object hence watermarking methodology preserves important distance relationships. Nearest Neighbors (NN) of each object and the Minimum Spanning Tree (MST) of the original dataset these are the algorithms used for dataset Preservation. It will leads to preservation of any mining operation that depends on the ordering of distances between objects and Preservation of Neighborhood property using Distance Relationship.

Keywords- Dataset, Watermarking, Nearest Neighbor, Minimum Spanning Tree.

I. INTRODUCTION

Data exchange and data publishing are becoming an inherent part of business and academic practices. Data owners also need to maintain the principal rights over the datasets that they share.This work presents a right-protection mechanism that can provide detectable evidence for the legal ownership of a shared dataset, without compromising its usability under a wide range of machine learning, mining, and search operations. This will be accomplished by guaranteeing that order relations between object distances remain unaltered. To right protect we use watermarking. Watermarking allows the user to hide innocuous pieces of information inside the data, adding descriptor to the Original Data so that whenever data gets Distributed over someplace by seeing the descriptor can find out from which source the data has taken. Here it will adapt on a robust spread-spectrum approach that can recover the embedded information even under malicious data transformations. Therefore goal is not only to provide right protection, but also to preserve important parts of the original object topology. It focus on the preservation of the following properties on the original distance graph: a) Preservation of Nearest-Neighbor (NN) distances for every object, and b) Preservation of the dataset's Minimum Spanning Tree (MST).

II. METHODS TO SOLVE THESE PROBLEMS

To Solve the Right Protection of dataset by using Distance Relationship we have to apply two MethodsBased on nearest neighborhood classification and Minimum Spanning Tree. Before applying Distance Relationship Methods first Use Watermarking Technique to insert the Watermark on dataset by using Fourier transformation then Extract it to get back original Dataset.

III. OVERVIEW OF WATERMARKING METHOD

Watermarking is meant by adding certain kind of descriptor to the original dataset, so that whenever data gets distributed over someplace by seeing the descriptor can make out who is the source. Assume an object represented as a vector of complex numbers of the form x = {x1…. xn}, Where $x_k = a_k + b_k i$(i is the imaginary unit, $i2 = -1$), and where the real and imaginary parts, ak and bk respectively, describe the coordinates of the kth point of object x on the imaginary plain. Apply the watermarking method on two dimensional dataset which can help to preserve the distance relationship. Assume the dataset of complex format and apply Fourier transformation method. In Fourier transformation can be classified as discrete Fourier transformation (DFT) and inverse of DFT i.e. IDFT, which are used for insertion and extraction of watermarking which will provide watermarked dataset and original dataset.

IV. DISTANCE DISTORTION DUE TO WATERMARK

When we apply Watermarking technique on dataset 'D' Using Fourier Descriptor which can give different coefficients and the robustness of the watermark embedding depends on the choice of coefficients. We embed the watermark in the coefficients that exhibit, on average over the dataset, the largest Fourier magnitudes. This makes the removal of the watermark difficult; in order for it to be masked out(e.g., by noise addition) it would mean that the dominant frequencies of the dataset have to be distorted.Assume a sequence$x \in C^n$ with corresponding set of Fourier descriptors X, and watermark $W \in R^n$ and power$P \in [0, 1]$ which specifies the intensity of the watermark. Amultiplicative watermark embedding (W, p) generates awatermarked sequence $\hat{x}$ by replacing the magnitudes of eachFourier descriptor of x with the watermarked magnitude $\hat{\delta}_j$while not altering the phases, specifically. Hence consider the watermark to be a given as $W \in \{-1, 0, +1\}^n$, which is embedded in all objects of the dataset. Hence by applying Watermarking method on dataset insertion and extraction of watermark on distance relationship to preserve the dataset property. Notations used D: Datasetδ: Magnitude, $\hat{\delta}$: Watermarked Magnitude, p: Power factor.

$\hat{\delta}_j = \delta_j . (1 + pW_j)$, and $\hat{\delta}_j = \delta_j$

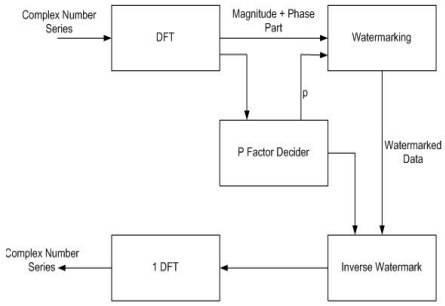

Figure 1. Overview of Watermarking Method.

Figure 1 Illustrates how watermarking method will work by using DFT and IDFT Method of Fourier Transformation. Hence embed the watermark in the magnitudes of the Fourier descriptors and leave the phases Unchanged and leave the DC component intact and watermark the Fourier descriptors with the largest average magnitudes. Here, study of analytically the distortion of object distances due to watermarking which provide closed-form expressions for the Euclidean distance between two objects before and after the watermark embedding. Then, derive tight lower and upper bounds on the distortion of distances due to watermarking, which hold uniformly for any watermark compatible with the given dataset. Hence Fourier transformation and distance relation methods are used in concept so the complex numbers are takes as input which can provide mapping the distance relation by its magnitude and phase value for datasets.

1) WATERMARK DETECTION: The detection process aims at discovering the presence of a particular watermark in a watermarked dataset. This involves measuring the correlation between a tested watermark and the watermarked dataset. The higher the correlation between the two, the higher the probability that the embedded watermark was the one tested. Because the watermark is embedded in all objects of a dataset, one option is to measure the correlation between the watermark and the average of the magnitudes of Fourier descriptors across all objects of the dataset. However, directly measuring the correlation may not be very effective under multiplicative embedding. The reason is that since we want to minimize distortion, a small embedding power is preferred, whence the magnitudes of the Fourier descriptors are dominated by the original level of the average. To overcome this, we record the bias μ(D) before embedding the watermark, and remove it before the detection. Denoting the vector of the averages of the Fourier descriptor magnitudes across all objects of a dataset D by $\mu(D) := (\mu_1(D), \ldots, \mu_n(D))$.

V. PROBLEM WITH DISTORTION

Distortion will affect the data mining algorithms which are distance based. When we run the data mining algorithm on original or watermarked data, it should give same results like same clusters etc. Since the amount of distortion is controlled by P value will can cause minimum impact on data mining algorithm by choosing a P value causing minimum distortion to distance functions. Lower and upper bounds of the distance expansion/contraction due to watermarking to show that two objects that lie at a given distance before the watermark embedding cannot get arbitrarily close or arbitrarily far apart after the watermark embedding, irrespective of the embedded watermark. In particular by using a restricted isometric property allowing the watermark embedding power to lie in the interval of $[p_{min}, p_{max}] \subseteq [0, 1]$, which will prove the minimum contraction factor to be $1 - p_{max}$ and the maximum expansion factor to be $1 + p_{max}$ for any dataset-compatible watermark. Important properties of the distance graph, because a number of mining, learning, and visualization algorithms are based on them. For example, preservation of the NN will result in preservation of search operations based on a query-by-example paradigm (e.g., multimedia search); instance-based classification tasks based on the Nearest-Neighbors will also be retained. Computation of the MST is also a fundamental operation in many data analysis tasks aMST of the distance graph in order to estimate theintrinsic data dimensionality. Using the 2D perimeter of a dataset one can easily visualize the evolutionary path between the different species. Our technique will guarantee that this based on the MST will produce same outputs, before and after watermarking. The algorithms that we put forward will properly tune the watermark embedding power p so that the NN of each object or the MST of the entire dataset are not distorted

There are two mechanisms

1).NN Preservation

2).MST Preservation

VI. NN PRESERVATION PROBLEM

Protection of watermarked datasets is the important object distances (NN or MST structures) are preserved.It is to optimize for maximal security, and detectability of the watermark, and at the same time Minimize visual distortion of objects. Therefore, we seek to find the maximum embedding power $p*$, so that the desired Properties are maintained.In certain cases, it may be possible to embed the watermark very strongly and still maintain the original distance relations, but this may lead to a visible distortion of an object's appearance.In practice we set an upper limit on the maximum allowed power, i.e., $p* \in [p_{min}, p_{max}]$. In our experiments we used value Pmax= 0.01, Allowing up to 1% relative distortion. This assures that Objects before and after watermarking will be virtuallyIndistinguishable. NN Watermarking problem here to find the nearest neighbour based on distance between the data points. To preserve the neighbourhood property on dataset NN watermark technique will be applied. It's difficult to find out least distortion among neighbourhood of dataset. Hence to find out least distortion we must calculate nearest neighbour with minimum 'p' value which can be taken as its least distorted point where the neighbourhood property will preserved.

VII. MST PRESERVATION PROBLEM

During MST Preservation technique the spanning tree structures must be preserved. MST preserving watermarking will allow for imperfect Preservation of the respective desirable property. In particular, it will allow for a few of objects to change their few of edges of the MST to be altered to provide higher security, one may be willing to accept introducing errors in the distance-based utility in exchange for a stronger watermark. Hence to show how to solve the respective problems: first, using exhaustive search (i.e., by computing and comparing all relevant distances between objects. MST Calculation is done through Prim-Kruskals algorithm which are used to find out least distortion. MST Tree Structure will not change even after

applying watermarking but to find out least distortion by using 'p' value for which we need to calculate the total distance by taking difference of MST before Watermarking and MST after Watermarking.

1) NN WATERMARK: NN Watermarking Method is used to provide the security over dataset for preservation of neighborhood property between the datasets. Hence watermark across multiple frequencies of each object and
Across multiple objects of the dataset will be applied. As such, it renders the removal of the watermark particularly difficult without substantially compromising the data utility. An object x is mapped into the frequency domain using its complex Fourier descriptors X = {X1. .Xn}. The mapping from the space domain to the frequency domain is described by the normalized discrete Fourier transform, DFT(x), and its inverse, IDFT(X) as described in Watermarking method. When we apply NN Watermark Method it will insert NN Watermark on a dataset and provides least distortion by using NN Power Finder with different 'p' values and distance between them will be calculated.
2) MST WATERMARK: MST Watermarking will be applied to preserve the tree structure after watermarking and get the minimum 'p' value for least distortion which can provide the tree structure with MST Watermark. Hence here also we need to apply Fourier transformation concept using DFT(x) and its inverse, IDFT(x) as used in Watermarking method. The MST-P Watermarking Problem can be solved again via a system of quadratic inequalities.

VIII. PERFORMANCE EVALUATION

Performance Evaluation of NN and MST is calculated based on its runtime as shown in the Figure 2. To evaluate the performance metrics we have to estimate the total CPU cost which can give computation of the coefficient of the quadratic terms. Number of quadratic inequalities which can give I/O cost
Of the algorithm. NN Performance is better than MST Performance because in NN Algorithm we are preserving Nearest Neighborhood property it will take less time compare to MST. Where as in MST it will take more time for preservation of tree structure. MST will build a tree structure which remains same even after watermarking.

NN and the MST, our formulation is applicable on any mining operation that depends on the order of distances between objects [1]. This makes our approach relevant for a wide range of distance-based learning, search, and mining algorithms. Hence visually demonstrate that the right-protection mechanism retains the desired parts of the underlying distance graph .In MST-Preservation algorithm. We use the visualization technique of [2], which uses the Minimum Spanning Tree to provide a lower dimensional projection of the original dataset and apply this Technique on the dataset and the resulting visualization Outcome and contrast the MST mapping of the original dataset against the MST of the watermarked dataset One can observe that the two MST's are almost identical for any practical purpose. Therefore here we confirm visually that the algorithm accurately chose the watermark embedding power so as to preserve the MST

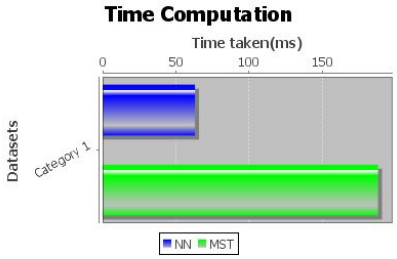

Figure 2. Performance graph of NN and MST

The above graph indicates the Performance metric of NN and MST which shows NN Performance will take less time compare to MST Performance.In addition, store the original distances of each object x from objects in its N(x)-neighbourhood, so that when computing squared watermarked distances, the constant term of the quadratic need not beRecomputed. The above interpretation highlights the reasoning for a potentially drastic reduction in time complexity attained by the Fast NN-Preservation algorithm: if the average size of each object's N-neighbourhood is O(1), the complexity of the algorithm after the pre-processing stage is no longer quadratic, but rather linear in the size of the dataset |D|. Because N can be empty for some objects, further reduction in the complexity is possible and the worst-case theoretical complexity is still Same.

IX. CONCLUSION

In the watermarking method on dataset which will provide right protection over dataset using distance relationship by preserving the properties of distance between datasets using algorithm NN and MST. Here it will calculate each neighborhood array and takes more time for computation of MST. Hence In future we can implement FAST MST and FAST NN algorithm which can calcite average of distance between dataset and increases the performance and speed of the algorithm.

REFERENCES

[1] IEEE Transactions on knowledge and data engineering, vol. 26, no. 8, august 2014.
[2] M. Vlachos, B. Taneri, E. J. Keogh, and P. S.Yu,"Visual exploration of genomic data," in Proc. 11th Eur. Conf. PKDD, vol. 4702,Warsaw, Poland, 2007, pp. 613–620.
[3] V. Solachidis and I. Pitas, "Watermarking polygonal lines using Fourier descriptors," IEEE Comput. Graph. Appl., vol. 24, no. 3 pp. 44–51, May/Jun. 2004.
[4] R. Agrawal and J. Kiernan, "Watermarking relational databases,"in Proc. 28th Int. Conf. VLDB, Hong Kong, China, 2002, pp. 155–166.
[5] Right-Protected Data Publishing with Hierarchical Clustering Preservation
[6] C. Lucchese, M. Vlachos, D. Rajan, and P. S. Yu, "Rights protectionof trajectory datasets with nearest-neighbor preservation," VLDBJ., vol. 19, no. 4, pp. 531–556, 2010.

